Introduction to Student Projects

IE500 Data Mining

UNIVERSITY OF MANNHEIM
Data and Web Science Group



Outline

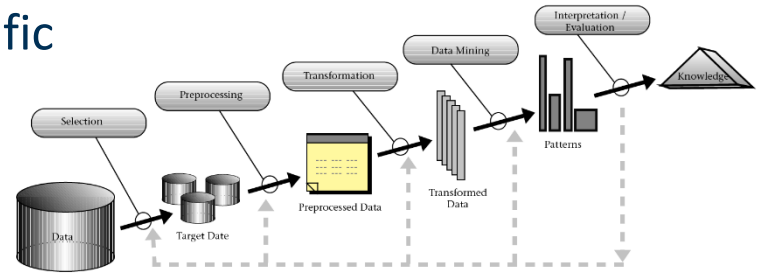
UNIVERSITY OF MANNHEIM
Data and Web Science Group

1. Requirements for the Student Projects
2. Requirements for the Project Reports
3. Final Exam
4. Team Formation

Student Projects

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **Goals**
  - Gain practical experience with the complete data mining process
  - Get to know additional problem-specific
    - preprocessing methods
    - data mining methods



- **Expectation**
  - You select an interesting data mining problem of your choice
  - You solve the problem using
    - the data mining methods that we have learned so far, including
      - proper hyperparameter optimization
      - problem-specific pre-processing and smart feature engineering
    - additional data mining methods which might be helpful for solving the problem and build on what we learned in class

Procedure

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Teams of **five (exceptionally also six)** students
  - realize a data mining project
  - write a 12-page summary of the project and the methods employed in the project
  - present the project results to the other students
    - 10 minutes presentation + 5 minutes discussion (might slightly be adjusted depending how many teams we will have)

Where to find interesting Data Sets?

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Data registries
  - Datasets hosted on Amazon AWS https://registry.opendata.aws
  - Google's Dataset Search: https://datasetsearch.research.google.com/
  - Microsoft Datasets: https://msropendata.com/
  - Yahoo Webscope Datasets: http://webscope.sandbox.yahoo.com/
  - Dataset collection on Github: https://github.com/awesomedata/awesome-public-datasets
  - Data Hub: http://datahub.io
  - Linked Open Data Cloud: http://lod-cloud.net/
  - Stanford Large Network Dataset Collection: http://snap.stanford.edu/data/index.html
  - Huggingface: https://huggingface.co/datasets

Where to find interesting Data Sets?

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Public sector data
  - US government: https://www.data.gov
  - UK government: https://data.gov.uk
  - EU: https://www.europeandataportal.eu
  - CIA World Fact Book: https://www.cia.gov/library/publications/the-world-factbook/
  - Health data (over 125 years): https://www.healthdata.gov/

Where to find interesting Data Sets?

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Competitions
  - Kaggle: https://www.kaggle.com/
  - Data Mining Cup: http://www.data-mining-cup.de
  - KDD Cup: https://www.kdd.org/kdd-cup
  - DrivenData: https://www.drivendata.org
  - CrowdAnalytix: https://www.crowdanalytix.com

- If you use a competitions task:
  You **have to** compare your results to results from the competition's forum!

Where to find interesting Data Sets?

UNIVERSITY OF MANNHEIM
Data and Web Science Group

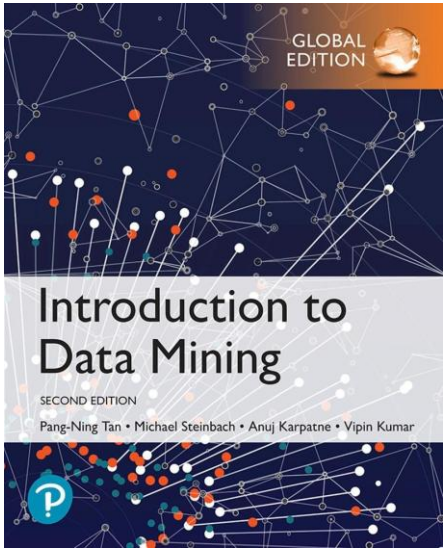
- Language resources
  - WordNet: https://wordnet.princeton.edu
  - EuroWordNet: http://projects.illc.uva.nl/EuroWordNet/
  - Project Gutenberg (36.000 ebooks): http://www.gutenberg.org/
  - New York Times (starts 1851): http://developer.nytimes.com/docs
  - Wikitionary: https://www.wiktionary.org
    as KG: http://kaiko.getalp.org/about-dbnary/

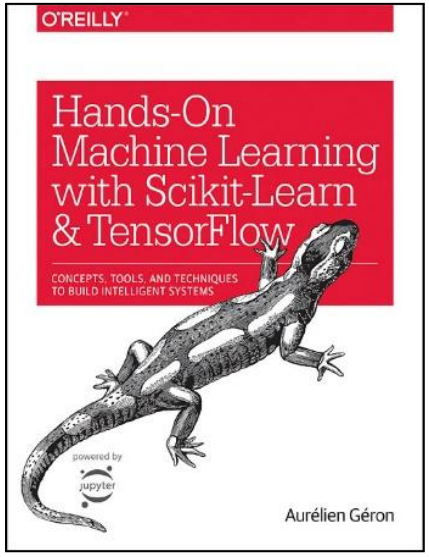
- Knowledge graphs
  - Wikidata: https://www.wikidata.org
  - BabelNet: https://babelnet.org
  - DBpedia: http://wiki.dbpedia.org

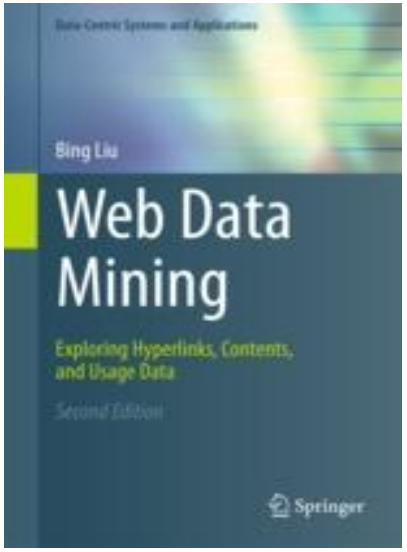
Where to Find Additional Information

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining, Pearson / Addison Wesley.
- Aurélien Géron: Hands-on Machine Learning with Scikit-Learn. O'Reilly.
- Bing Liu: Web Data Mining, 2nd Edition, Springer.







Where to Find Additional Information

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Check out the solutions to your problem that other people have tried.
  - by looking into the Kaggle discussion groups and code
  - by investigating the state-of-the-art for your your task on Papers with Code (https://paperswithcode.co/)
  - by looking at submissions of the KDD Cup or Data Mining Cup
  - or search for relevant scientific papers using Google Scholar, search term:
    “task name + survey”

Papers With Code

kaggle

Google

CHAT GPT

DATA MINING CUP
International Student Competition

Some Projects realized in previous Semesters

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Airbnb (done very often)
  - predict the prices of new apartments
  - E.g. Predicting HDB Resale Flat Prices in Singapore
- Spotify Music Popularity Classification (done very often)
  - Classify how popular a song will be based on features already known when the song is submitted
- Bundesliga Betting Rules
  - Find rules that help you to predict the outcome of a Bundesliga game
- Predicting Urban Bicycle Flow in Mannheim
  - When and where are bicyles needed; how many bikes cross Kurpfalzbrücke in the next hour
- Predicting Injury Severity in Road Accidents
- Flight price prediction

Some Projects realized in previous Semesters

- Airbnb (done very often)
  - predict the prices of new apartments
  - E.g. Predicting HDB Resale Flat Prices in Singapore
- Spotify Music Popularity Classification (done very often)
  - Classify how popular a song will be based on features already known when the song is su
- Bunde
  - Find rules that help a game
- Predicting Urban Bicycle Flow in Mannheim
  - When and where are bicyles needed; how many bikes cross Kurpfalzbrücke in the next hour
- Predicting Injury Severity in Road Accidents
- Flight price prediction

Choose a task/dataset where you have a ground truth (or can easily generate one)

Dataset Selection: Key Considerations - Pros

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **Rich Feature Space**
  - Datasets should have multiple, diverse features that allow for creative feature engineering
- **Adequate Sample Size**
  - Aim for datasets with at least 10,000 examples to ensure robust modeling
- **Balanced Complexity**
  - A dataset should be complex enough to challenge students without being computationally prohibitive.
- **High Data Quality**
  - Ensure key columns are well-populated (e.g., <5% missing values) so that the data can be effectively used
- **Novelty**
  - Prefer datasets that haven't been overused in existing challenges, offering room for innovative approaches

Dataset Selection: Key Considerations - Cons

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **Overly Simple**
  - Avoid datasets with too few features (< 5) or a too-basic topic, as this limits feature engineering.
- **Excessively Large**
  - Datasets with over 1 million records (e.g., huge product datasets) can be too compute-intensive
- **Over-Saturated**
  - Datasets with clear guidelines and abundant available code (e.g., well-established challenges)
- **Poor Data Usability**
  - Be wary of datasets where important columns are empty more than 5% of the time, or where the ground truth is ambiguous
- **Additional Tip**
  - Check prior usage—if you're the first to work on a dataset, verify that the dataset is practically usable and that data quality issues won't undermine your project

Team Formation

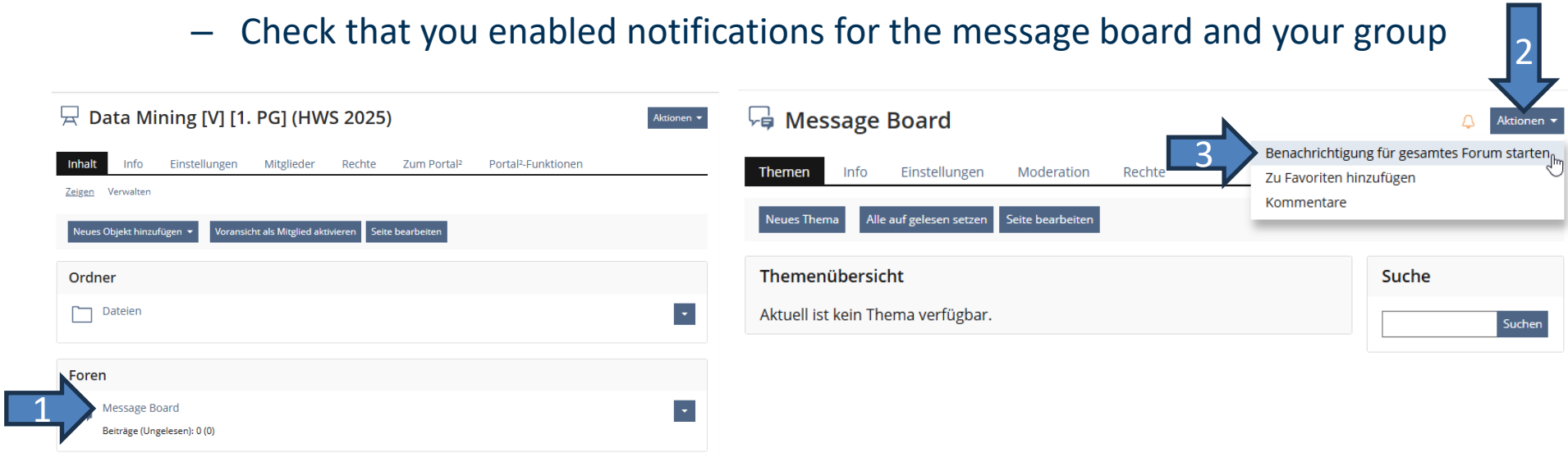
UNIVERSITY OF MANNHEIM
Data and Web Science Group

- You are allowed to form teams of **five (exceptionally also six)** students as you like!
  - You enter your team into the Group Formation spreadsheet (see last slide) until **Monday, October 5th 23:59**
  - If you are less than five you can still enter your team (but you will be assigned new team members)
  - If you are still looking for a team, use the Ilias message board to find teams
    - If you do not find a team, also write in the Ilias message board that you are still searching for a team and we will assign you to a team shortly after the deadline



Team Formation

- Once your team is formed:
  - Meet with your team to organize your work
    - Decide project topic
    - Organize writing of project outline
  - You can start writing the project outline
- For further communication, we will create groups in Ilias
  - Check that you enabled notifications for the message board and your group



Team Formation

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Project and Exam are two parts of the module
  - Once you are registered for the team project, you also **have to** register for the exam
  - Once the teams are formed, we will inform the student office about your registration

Project Outlines

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Maximum 4 pages (sharp!) including title page
  - Using DWS master thesis layout (PDF!)
  - Include on the first page:
    - **project name**
    - **your team number and**
    - **all names**
- Due **Wednesday, October, 14th, 23:59 (Submission via Ilias)**
- By **Friday, October 16th:** Feedback on your proposal
  - If revisions are required, discuss questions with **Franz via your team's ILIAS forum**
- Revised proposal due **Monday, October 26th, 23:59**
- Feedback on revised proposals by **Monday, November 2nd**

Project Outlines

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Answer the following questions:
  1. What is the problem you are solving?
  2. What data will you use?
     - Where will you get it?
     - How will you gather it?
  3. How will you solve the problem?
     - What preprocessing steps will be required?
     - Which algorithms do you plan to use? Be as specific as you can!
  4. How will you measure success? (Evaluation method)
  5. What do you expect your results to look like? (Model/Clusters/Patterns)

Coaching Sessions

UNIVERSITY OF MANNHEIM
Data and Web Science Group

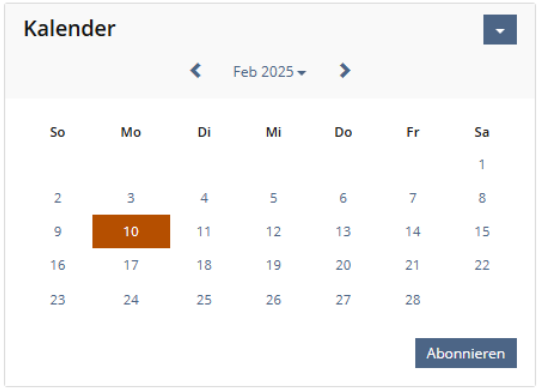
- We will give you tips and answer questions concerning your project
- At the time of the lecture (Mondays)

- **Every team has to attend at least one coaching session!**
  - Attending multiple times is allowed and recommended

Course Organization - Calendar

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- This semester, we will **use the calendar feature in Ilias** to schedule coaching sessions.
  - You register via the calender in Ilias for your coaching session
  - A few days before your session, we will **inform you via the main forum** about your **exact time slot** within the 90 min window.
  - **Only one person per group** should book a slot on behalf of the group.
  - When booking, you must include your **group number** and a few **questions or topics** you want to discuss. **Blank requests will be ignored!**
  - The registration opens one week before it (Monday, 16:00)



Some Project Management Hints

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Organize your project in **multiple iterations**
  - Every artefact will be improved over time!
- Get a **simple process running early** on to have a baseline
- **Parallelize tasks** while keeping centrally track of results
  - e.g. one central document with results plus reference to exact version of the notebooks/datasets that produced these results
  - sub-groups should explore specific ideas for a specified amount of time

Some Project Management Hints

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **Define concrete milestones**: When should what be finished?
  - e.g. 05.11.26 Data exploration results collected in single document
  - e.g. 10.11.26 Subgroup on sentiment lexica adds results to central document

- **Infrastructure**
  - use shared folder for result document, versions of data, processes, slideset (e.g. MS Teams, Google Drive, github)
  - use LLMs for inspiration about additional methods as well as coding

Tasks within the Iterations of the Project

UNIVERSITY OF MANNHEIM
Data and Web Science Group

1. Data Exploration and Visualization
2. Data Preprocessing: value normalization, deal with outliers, deal with missing values, feature generation, balance training data if necessary
3. Establish/update baseline (majority class, predict mean value)
4. Try different learning methods using different feature creation methods and feature combinations
5. Perform error analysis in order to understand what is going on!
6. Later iteration:
   - run automatic hyperparameter optimization and attribute selection
   - employ more sophisticated evaluation setup: x-val + holdout vs. nested x-val

Project Report

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Max. 12 pages including title/toc page and reference page
  - max. 10 pages content, no appendix
  - Each extra page and each day of late submission downgrades your mark by 0.3!
- Reports and additional material need to be uploaded in Ilias within the respective Ilias groups
  - **Deadline: Sunday, December 6th, 23:59**

Project Report

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Outline for project report:
  - Application area and goals (0.5 pages)
  - Profile (structure and size) of your data set (minimum 1 page)
  - Preprocessing
  - Data Mining
    - Describe different approaches and parameter settings/optimizations that you tried
  - Evaluation
    - Including description of evaluation setup (split, x-val, nested-x-val?)
    - Including an analysis of the errors still made by the best method, a discussion of the results, and a comparison to state-of-the-art results (together: minimum 2 pages)
  - Results

Project Report

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Requirements
  - You have to use the latex template of the DWS Thesis
  - Please cite sources properly and use your references page
  - Also submit your Python code and (a subset) of your data
  - Include your names and your team number on the first page!

- Usage of AI Tools needs to be declared

Declaration of Used AI Tools

Tool	Purpose	Where?	Useful?
ChatGPT	Rephrasing	Throughout	+
DeepL	Translation	Throughout	+
ResearchGPT	Summarization of related work	Sec. 2.2	-
Dall-E	Image generation	Figs. 2, 3	++
GPT-4	Code generation	functions.py	+
ChatGPT	Related work hallucination	Most of bibliography	++

Checklist for Project Reports

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Business Understanding
  - What is the actual problem (in the domain)?
  - What is the target variable?
    - Classification/Regression/Cluster Analysis?

- Data Understanding
  - What is the distribution of labels / target variable?
  - Are all attributes and their types listed and important attributes explained?
  - What is the quality of the data? Wrong values? Outdated?
  - What does correlation analysis reveal about attribute importance?

Checklist for Project Reports

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Preprocessing
  - Are missing values replaced (in case needed)?
  - Checked for outliers (and handled them)?
  - Validity tests of attributes (Height above sea level < 9000)?
  - Check for inconsistencies (age=42, birthday=03/07/1997)
  - Check for duplicates
  - Performed data normalization (e.g. US vs United States)
  - Additional features generated?
  - Has binning been tried out?
  - Feature subset selection necessary?
- External Knowledge:
  - Are additional datasets used?

Checklist for Project Reports

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- ML approaches
  - How many different ML approaches were tried out?
  - Do you have at least one symbolic and one non symbolic approach?
  - Do you have at least one baseline (majority class / mean value / domain specific …)?

- Evaluation
  - Is there a train test split or 10-fold cross validation implemented
  - Is the evaluation stratified?
  - Cost matrix or not?
  - Are the hyper parameters tuned (in which range / which attributes) ?
  - Are the tests systematic?
  - Analyse a symbolic model (how does the decision tree / rules /… looks like)
  - What features do have a high impact on the result?

Checklist for Project Reports

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Result
  - Is the result critically evaluated
  - Is the result analyzed against the baseline
  - What does the result mean given the problem (could you use it)

Project Presentation

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Present the project results to the other students
  - 10 minutes presentation + 5 minutes discussion
  - During exercise slot

- Presentations need to be uploaded in Ilias within the respective Ilias groups (as PDF)
  - **Deadline: Wednesday, December 9th, 23:59**
- Three **90-minute sessions** will be available.
- For **presentations, attendance is mandatory per session** for all group members, so the exact timing within the session does not matter.
  - You can decide if all members present or only a few
  - Questions need to be answered by each of the members
- Keep an eye on the **general forum**—we will announce the exact time when slots become available **at least one week in advance**.

Get Additional Advice from a Stanford Professor

UNIVERSITY OF MANNHEIM
Data and Web Science Group



Christopher Potts

- How to evaluate your model?
  - https://www.youtube.com/watch?v=TxTblROT9lY
- How to structure your project report?
  - https://www.youtube.com/watch?v=DZNwO-p5PGY
- How to present the results of your project?
  - https://www.youtube.com/watch?v=GGx7klcahzY

Deadlines - Overview

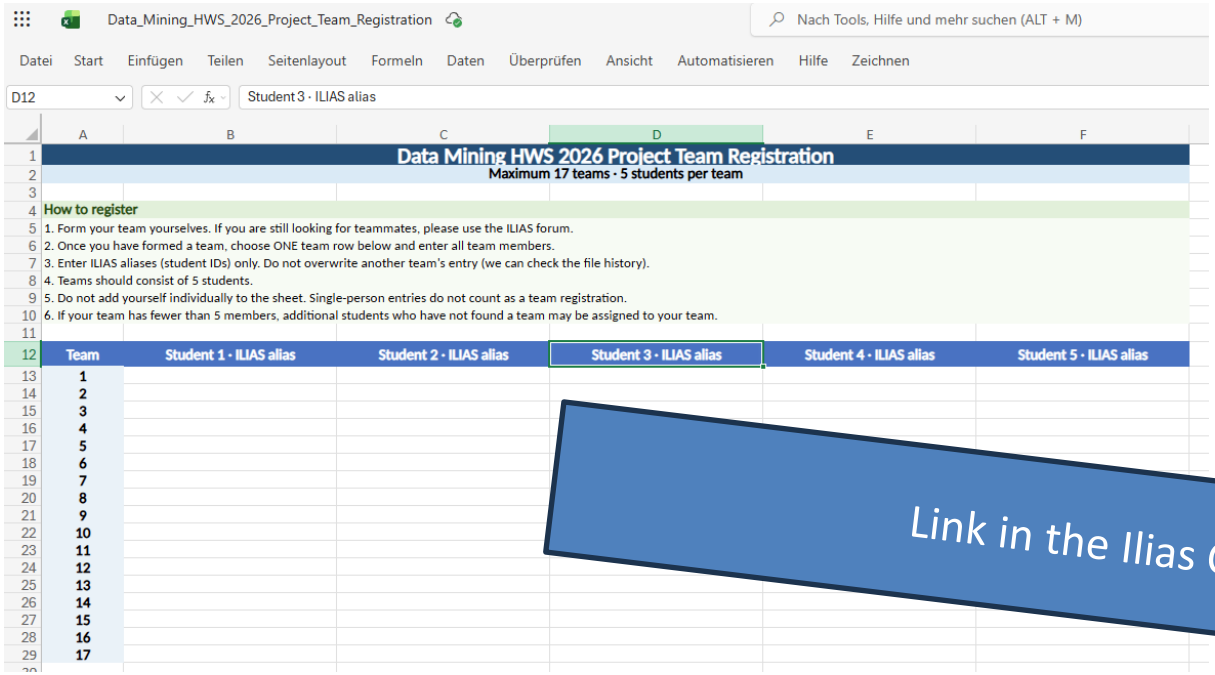
UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Team formation until **Monday, October 5th 23:59**
- Project outline until **Wednesday, October, 14th, 23:59**
- Coaching Sessions
  - Every team has to attend at least one coaching session
- Project report until **Sunday, December 6th, 23:59**
- Project presentation in PDF until **Wednesday, December 9th, 23:59**

Team Assignment

- Enter your group in “Team Setup” in Google Sheet
  - In case you do not have a team, fill in your details in “Looking for a team”

  => then you will be assigned to a team after the registration period
- Do so until Sunday October 5th 23:59



Link in the Ilias Course

Questions?

UNIVERSITY OF MANNHEIM
Data and Web Science Group

